FERMILAB-CONF-04-492-CD

Testing the CDF Distributed Computing Framework

Valeria Bartsch[2], Andrew Baranovski[1], Stefano Belforte[3], Mòrag Burgon-Lyon[4], Gabriele Garzoglio[1], Randolph Herber[1], Robert Illingworth[1], Rob Kennedy[1], Ulrich Kerzel[5], Art Kreymer[1], Matt Leslie[2], Lauri Loebel-Carpenter[1], Lee Lueking[1], Adam Lyon[1], Wyatt Merritt[1], Fedor Ratnikov[6], Alan Sill[7], Richard St. Denis[4], Stefan Stonjek[1,2], Igor Terekhov[1], Julie Trumbo[1], Sinisa Veseli[1] Steve White[1]
[1]Fermi National Accelerator Laboratory, [2]University of Oxford, [3]INFN / Trieste
[4]University of Glasgow [5]Universität Karlsruhe [6]Rutgers University [7]Texas Tech University

Abstract

A major source of CPU power for CDF (Collider Detector at Fermilab) is the CAF (Central Analysis Farm) [1] at Fermilab. The CAF is a farm of computers running Linux with access to the CDF data handling system and databases to allow CDF collaborators to run batch analysis jobs. Beside providing CPU power it has a good monitoring tool. The CAF software is a wrapper around a batch system, either fbsng [3] or condor, to submit jobs in a uniform way. So the submission to the CAF clusters inside and outside Fermilab from many computers with kerberos authentification is possible. It is mainly used to access datasets which comprise a large amount of files and analyze the data. Up to now the DCache system has been used to access the files. In autumn 2004 some of the important datasets will only be readable with the help of the data handling system SAM (Sequential Access to data via Metadata) [2]. This will be done in order to switch to using only one data handling system at Fermilab and on the remote sites. SAM has been used in run II to store, manage, deliver and track the processing of all data. It is designed to copy data to remote sites with remote analysis in mind. To prove CAF and SAM could provide the required CPU power and Data Handling, stress tests of the combined system were carried out.

A second goal of CDF is to distribute computing. In 2005 50% of the computing shall be located outside of Fermilab. For this purpose CDF will use the DCAF (Decentralized CDF Analysis Farms) in combination with SAM. To achieve user friendliness the SAM station environment has to be common to all stations and adaptations to the environment have to be made.

SYSTEM ANALYSIS

The CAF system has about 800 nodes with a total CPU of 1200 GHz, the computing power of the CAF will of course increase with time. In addition to that the Condor-CAF, a CAF system based on the batch system Condor has a total CPU of 2000 GHz and around 400 nodes. As can be seen in figure 1 at the moment 25 TBytes per day are read by the CDF community. Most of these data are read at the CAF at Fermilab with the DCache System [4]. The most important data are on disk (around 5 TB) and about 1 PB are on tape.

Testing the SAM environment therefore means to create

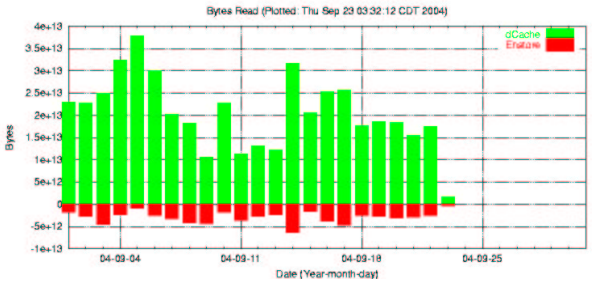

Figure 1: Bytes read in September 2004 from dCache at Fermilab by the CDF community in green and Bytes stored in red. In total files amounting about 25 TBytes are read on a daily basis.

the usual user load on the CAF system. In other words to load the CAF with about 50 SAM projects and move around 20 TBytes per day. The SAM projects should be splitted into several segments to run several parts of the jobs in parallel on different CPUs.

At the time of writing the DCAF systems excluding the original CAF at Fermilab and the CondorCaf at Fermilab total CPU of about 2000 GHz and 40 TBytes of disk space increasing with time. In contrast to the central CAF system the data transfer rates are quite small (depending on the network in the order of 500 GBytes per day), therefore SAM jobs submitted to the DCAF systems need to read the data already transferred to the systems. Testing the DCAF systems should make sure that a unique interface to the user has to be provided. That means environment variables like the name of the SAM station and the project name be set automatically. In addition the data transfer to the DCAF SAM station has been tested.

TESTS

Figure 2 shows the total size of all files transferred with the data handling system SAM grouped by SAM stations. The largest amount of data has been transferred to the central SAM station at Fermilab, cdf-sam. The location of a file is denoted by a URL to a file. In total the URL's of files amounting 600 TBytes have been consumed on the central SAM station at Fermilab. Not all of these files have been actually transferred and read but only the URL checked. On off-site SAM stations all files needed to be transferred inde-

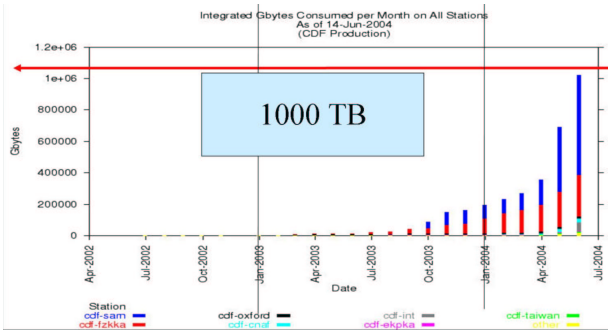

Figure 2: Total size of consumed files per month on all SAM stations since 2002 until 2004. In total about one PBytes of data have been transferred for CDF. The largest number of files were consumed by the central SAM station at Fermilab, cdf-sam. Not all of these files have actually been read, for some of them only the location has been checked. The local SAM station at Karlsruhe, Germany, cdf-fzkka, has also consumed a large amount of files. The data were used by the local CDF users at Karlsruhe. The DCAF SAM stations, cdf-cnaf and cdf-taiwan to name two of them, have transferred about 46 TBytes since the beginning of this year.

pendent of the readout program, but not all of them might have been read and analysed. The second largest amount of data has been transferred to the SAM station at Karlsruhe, Germany, cdf-fzkka, where local CDF users work with SAM as their only data handling system for analysis. At the SAM stations connected to the DCAF systems in total 46 TBytes have been transferred since their installation this winter. The administrators of the DCAF systems have imported datasets interesting for their users. A list of the datasets is published and automatically updated [5].

Figure 3 shows that the typical user load of about 25 TBytes of data each day has been created during the stress tests. While submitting jobs with several segments different DCache doors were assigned to the project. A DCache door is a network server which performs user authentication and forwards client requests to the pool managers. The assigment to more than one DCache door had an effect on the end of the projects. Because the signal to stop the project was sent when all files of the segment were read, but this was not necessary equal to the time when all segments have read all files. The problem was solved by assigning only one DCache door to one SAM job.

A recovery command of partly failed projects has been developed. This command depends on the correct release of files and can handle projects with several segments.

EXPERIENCED AND RESOLVED ISSUES

When submitting more than 100 SAM projects at one time problems with the project master have been observed.

SAM projects requesting a dataset with a large number of files (in this example 27000 files) also cause problems.

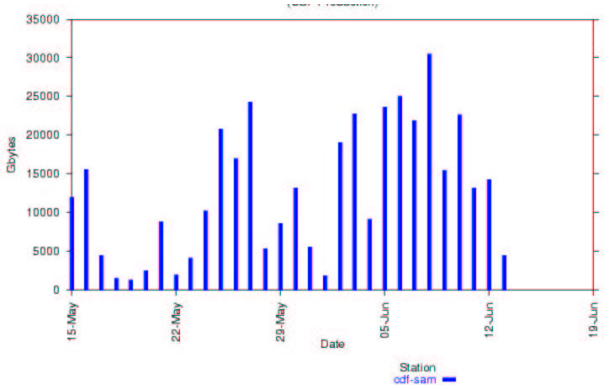

Figure 3: As an example the amount of data transferred during one of the test phases is shown. The typical load of 25 TBytes of data read on a daily basis has been created by test jobs on the central SAM station at Fermilab, cdf-sam.

The optimizer had to be improved to prevent checks on the location of all files when requesting the next file of a dataset or after a network outage. Otherwise the optimizer would be blocked and therefore on all SAM stations the file retrieval would slow down. This issue has been resolved in the latest version of the optimizer.

During the stress tests the load on the SAM station and the database server has been monitored and was most of the time quite low. So on this site one will not run into problems.

CONCLUSION

The tests of SAM on the CAF system have shown that SAM can be deployed by physics users this autumn. Some problems have been solved and a recovery program has been developed to analyze failures and recover the project. A second series of testing has to be done in order to check the limits of the data handling system SAM. The DCAF systems can be used when restricting the user to data which have already been imported to the decentralized SAM stations.

REFERENCES

[1] http://www.cdfcaf.fnal.gov

[2] http://d0db-prd.fnal.gov/sam/

[3] CHEP 01, Bejing, FBSNG - Batch System for Farm Architecture, J. Fromm et al.

[4] http://dcache.desy.de/

[5] http://hexfm1.rutgers.edu/DATA_INFO/sam_data/